VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

SYSTÉM PRO SPRÁVU A MONITORING WEBŮ

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

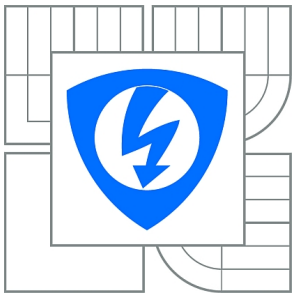
AUTOR PRÁCE
AUTHOR

MATĚJ POLÁCH

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF TELECOMMUNICATIONS

SYSTÉM PRO SPRÁVU A MONITORING WEBŮ

SYSTEM FOR WEBSITE MANAGEMENT AND MONITORING

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

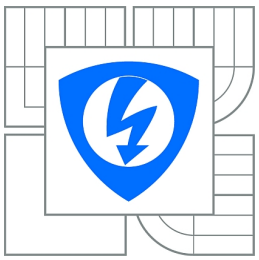
AUTOR PRÁCE — MATĚJ POLÁCH

AUTHOR

VEDOUCÍ PRÁCE — Ing. MARTIN KOUTNÝ, Ph.D.

SUPERVISOR

BRNO 2013



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Bakalářská práce

bakalářský studijní obor
Teleinformatika

Student:	Matěj Polách	***ID:***	125599
Ročník:	3	***Akademický rok:***	2012/2013

NÁZEV TÉMATU:

Systém pro správu a monitoring webů

POKYNY PRO VYPRACOVÁNÍ:

Realizujte systém pro správu a monitoring webů. V systému bude možné zaznamenat stránky, které nás zajímají, a sledovat u nich parametry v čase s grafickým a textovým reportem. Mezi tyto parametry patří např. page rank, srank, počet návštěv, počty interních a externích odkazů, počty facebook přátel aj.. V projektu realizujte nebo nainstalujte systém pro správu zpětných odkazů a integrujte jej do realizovaného prostředí. Další vlastnosti a schopnosti systému budou upřesněny v průběhu realizace.

DOPORUČENÁ LITERATURA:

[1] Nette framework [online]. 2013 [cit. 2013-02-10]. Dostupné z: http://nette.org/cs/

Termín zadání:	11.2.2013	***Termín odevzdání:***	5.6.2013

Vedoucí práce: Ing. Martin Koutný, Ph.D.
Konzultanti bakalářské práce:

prof. Ing. Kamil Vrba, CSc.
Předseda oborové rady

ABSTRAKT

Tato bakalářská práce realizuje systém pro správu a monitoring webů. Práce srovnává a vyhodnocuje současné možnosti monitorování webů. Dále popisuje vlastní řešení aplikace a její možnosti. Vytvořený systém zaznamenává stránky, které zajímají jejich správce a sleduje u nich parametry v čase. Významnými vlastnostmi systému jsou schopnost sledovat stránky bez nutnosti zasahovat do jejich zdrojového kódu (tedy i konkurenční) a možnost sledovat daná kritéria v průběhu času. Systém poskytuje výstupy v textové a grafické podobě. Je založen na svobodných technologiích.

KLÍČOVÁ SLOVA

PHP, Nette, web, systém, SEO, monitorování

ABSTRACT

This bachelor thesis implements system of web management and monitoring. The thesis compares and evaluates contemporary possibilities of web monitoring. It likewise describes the application solution and its capabilities. Created system records the managed webpages, monitors their parameters during the time. The system has significant features: Ability to monitor webpages without source code treatment (possibility to monitor competitive webpages as well) and tracking the criteria within the time. The system provides outputs in both text and graphic shapes. It is based on free technologies.

KEYWORDS

PHP, Nette, web, system, SEO, monitoring

POLÁCH, Matěj *Systém pro správu a monitoring webů*: bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2013. 29 s. Vedoucí práce byl Ing. Martin Koutný, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Systém pro správu a monitoring webů“ jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

(podpis autora)

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu bakalářské práce Ing. Martinu Koutnému, Ph.D. za odborné vedení, konzultace a cenné podměty. Dále také tvůrcům a komunitě použitých softwarových produktů a mé rodině za podporu a zázemí, bez nichž by tato práce nebyla realizovatelná.

Brno ....

(podpis autora)

Obsah

Seznam obrázků

Úvod

V současné době se stalo každodenní samozřejmostí využívání nespočtu internetových služeb (web, email, voip). Pro běžné uživatele je internet nejčastěji zdrojem stránek, které si zobrazuje na svém zařízení a využívá jejich informační, vzdělávací, uměleckou a podnikatelskou hodnotu. Takový uživatel většinou nezná (ani to nepotřebuje) pozadí a způsob fungování webové služby. Požaduje většinou jen snadný a rychlý přístup ke svému oblíbenému obsahu.

Obsah webu vytváří a udržuje další kategorie jeho uživatelů – amatérští i profesionální tvůrci, kutilové i webdesignéři, kodéři a administrátoři. Jejich zájmem a požadavkem je zaujmout a přitáhnout na své stránky co nejvíce návštěvníků. V dnešní hojnosti nabízeného obsahu je to velmi náročný úkol a k tomu, aby mohli svou práci efektivně zlepšovat, se neobejdou beze zpětné vazby. Vedle základních statistických údajů o přístupu ke svým stránkám potřebují sledovat i řadu dalších charakteristik ukazujících na chování návštěvníků, uživatelů a přijímat opatření k udržení věrného zájmu a častých návštěv. Toto chování sice lépe vystihnou systémy, které mají přístup ke stránkám zevnitř (vyžadují speciální kód vložený ve stránce), ale existují i jiné způsoby získávání údajů o stránce. Tomuto se věnuji v následujících kapitolách.

Aplikace, která je předmětem této práce, umožňuje sledovat, zaznamenávat a zpracovávat různé parametry webů a jejich změny podle toho, které moduly si uživatel u daného webu zvolí. Aplikace umí pracovat se všemi weby (statickými i dynamicky vytvářenými) nezávisle na rozličnosti použitých technologií a redakčních systémů.

Pro realizaci systému jsem zvolil formu webové aplikace. Možnosti jako desktopovou aplikaci, aplikaci pro mobilní telefony a tablety jsem se rozhodl nepoužít.

Systém, který je předmětem této práce, vypadá navenek jednoduše a nabízí řadu užitečných nástrojů a pomůcek pro správce webů. Jeho předností oproti podobným službám však je, že nevyžaduje žádné zásahy do obsahu sledovaných stránek – umožňuje tedy například i monitorování konkurence. Pracuje rovněž na síti a využívá její rozšířené a oblíbené prostředky a nástroje (Linux, Apache, PHP, MySQL a Nette).

1 Řešení studentské práce

1.1 Současné možnosti monitorování webů

1.1.1 SEO – optimalizace pro vyhledávače

SEO je významná činnost, která slouží ke zvýšení pravděpodobnosti vyhledání požadované webové stránky. Její důležitost oceňují zejména zejména pracovníci marketingu a obchodu. Správná optimalizace webu pro vyhledávače může významně ovlivnit ochodní úspěch firmy.

Jennifer Grappone a Gradiva Couzin v [2] (s. 12) píší, že pro vyhledání cílového publika vašeho webu, tedy těch, kdo se snaží najít vaši společnost, váš produkt, vaši službu nebo ten druh informací, který se nachází na vašem webu, je potřeba vypátrání těchto lidí a následné vytvoření mimořádně zacílené informace, kterou umístíte tam, kde ji snadno naleznou.

Stejné autorky vyjmenovávají řadu přesvědčivých důvodů k tomu, aby se autoři stránek zabývali problematikou SEO. Docházejí zároveň k závěru, že SEO vyžaduje hodně vytrvalosti.

Tato skutečnost je pro tuto práci výchozí jako odůvodnění požadavku, aby nástroje a postupy, které jsou cílem této diplomové práce, poskytovaly možnost dlouhodobého sledování, porovnávání a vyhodnocování sledovaných kritérií. Vývoj a trendy sledovaných klíčových hodnot jsou využitelné pro analýzu procesu a pro přijímání opatření k nápravě a posléze k ověřování účinnosti těchto opatření (zpětné vazbě) ([2] od s. 34).

1.1.2 Monitorování konkurence

V oblasti SEO je významným prvkem monitorování konkurence. [2] od s. 138 věnuje této problematice významný prostor. Popisuje taky způsoby, možnosti a nástroje, jak monitorování provádět, které jsou rozepsány níže. Tyto informace mě vedly ve své práci k tomu, aby cíle mé práce – vytvářené nástroje – umožňovaly i sledování a porovnávání konkurence.

1.1.3 Prostředky a metody monitorování webů

V současné době existuje velké množství prostředků a metod použitelných pro monitorování webů. Na některé z těchto služeb odkazuje web `tvorba-webu` [3] a popisuje jejich možnosti. Funkčnost těchto prostředků a metod je rozmanitá. Rozdílné funkční možnosti a omezení těchto nástrojů určují vhodnost použití k určitým účelům.

Následuje stručný rozbor výběru dostupných prostředků a hodnocení jejich vhodnosti pro ten který účel.

Ruční metoda hodnocení

Nejdostupnějším nástrojem pro shromažďování a vyhodnocování informací je ruční metoda. Ta může využívat nástroje jako např. tabulkový procesor (MS Excel nebo OpenOffice Calc). Nevýhodou takového řešení je zdlouhavé pořizování a zpracování dat. Data nelze aktualizovat snadným, rychlým a automatickým způsobem. Takový nástroj je příliš závislý na lidském faktoru.

Ruční kontrola je však nezastupitelná například pro prohlížení a srovnávání výsledků získaných z vyhledávačů. [2], s. 133 pro ni uvádí tři dobré důvody:

1. Ruční kontrola je přesnější než automatická.
2. Ruční kontrola vás bude udržovat v blízkém kontaktu se změnami v hodnocení.
3. Většina vyhledávačů se na automatizovanou kontrolu hodnocení dívá s velkou nelibostí, protože tyto nástroje v krátkém čase odesílají spoustu dotazů. To zatěžuje vyhledávač.

Zdrojový kód stránky

Prohlížení zdrojového kódu stránky je snadno dostupným nástrojem [2], s. 141. Je vhodný pro „ruční“ prohlížení obsahu tagů jako `META` (metadata [1] jako např. důležitá `KEYWORDS` nebo `TITLE`). Výhodou je, že jej lze použít na jakýchkoliv stránkách – i konkurenčních. Nevýhodou je, že není snadné a pohodlné používat tento způsob opakovaně.

Možnost číst „neviditelný“ obsah webových stránek však poskytnou některé automatizované nástroje.

Google Toolbar

Google Toolbar [2], s. 140 je nástroj, který lze doinstalovat do běžných webových prohlížečů a který umožňuje sledovat hodnotu Page Rank [2] [2], s. 312 a backward links [3] (= zpětné odkazy).

K Page Ranku a dalším informacím je rovněž možno dostat se prostřednictvím služby na adrese `www.faganfinder.com/urlinfo/`.

[1] Skryté informace, které specifikují různé vlastnosti dokumentu (například autora dokumentu, strukturu nebo klíčová slova). Metadata mohou být ve vyhledávačích použita pro určení relevance a pozice. [2], s. 311

[2] Googlem patentovaný algoritmus pro určování důležitosti webové prezentace. Hodnota PageRanku se pohybuje od 0 do 10, kde 10 je nejvyšší úroveň důležitosti. PageRank je občas označován pomocí zkratky PR.

[3] Odkazy, které na naše webové stránky směřují z jiných webových stránek, Backlink

Alexa

Alexa ([2], s. 142) je databáze umístěná na `www.alexa.com`. Poskytuje následující informace o webových stránkách:

1. Screenshot domovské stránky
2. Data o návštěvnosti
3. Zpětné odkazy
4. Kontaktní informace o majiteli
5. Odkaz na staré verze webu z archivu
6. Počítá hodnocení návštěvnosti vůči ostatním stránkám na webu.

Alexa poskytuje i nástrojovou lištu do prohlížeče (podobně jak Google Toolbar).

Google Analytics

Jedná se o oblíbený a rozšířený pokročilý systém pro monitorování webových stránek ([2], s. 218). Sleduje a vyhodnocuje návštěvnost v průběhu času a je schopen rozlišit návštěvníky např. podle geolokace či technických vlastností koncového zařízení (rozlišení displeje). Sleduje chování návštěvníků (průměrnou délku návštěvy, způsob vstupu na stránku). Poskytuje číselné a grafické interaktivní výstupy. Systém analyzuje inzerci. Systém je k dispozici zdarma. Funkčnost systému je však podmíněná vložením kódu do sledovaných webových stránek.

Toplist.cz

Nástroj Toplist umožňuje vložení infopanelu na své stránky, tabulky obsahující údaje o návštěvnosti celkové, týdenní, dnešní a aktuální. Dále zobrazuje umístění webu ve svém žebříčku. Zmíněné informace jsou na stránku vkládány formou obrázku, což komplikuje další využívání číselných údajů. Pro účel odpovídající zadání této práce je prakticky nevyužitelný,

Navrcholu.cz

Služba Navrcholu (Internet Info) je nástrojem sledování návštěvnosti webových stránek. Její fungování je podmíněno vložením kódu do sledované stránky, který odkazuje na objekt poskytovatele služby. Ten získává potřebné údaje, zpracovává je a poskytuje zaregistrovanému uživateli.

Awstats

Nástroj Awstats je analyzátor webových logů[4]([2], s. 309), skript fungující na serveru poskytovatele sledovaného webu a analyzuje přístupový a chybový záznam serveru (`access.log` a `error.log`). Jeho výstupy jsou číselné i grafické a umožňují sledovat návštěvnost a chování uživatelů. Použití tohoto nástroje je omezeno pouze na oprávněného uživatele.

Všechny výše uvedené nástroje mají jednu společnou nežádoucí vlastnost: Je nutný zásah do zdrojového kódu sledovaných stránek nebo nastavení na straně serveru. Tím je vyloučeno sledování a porovnávání webu konkurence.

Validátory kódu

Slouží ke kontrole správnosti zdrojového kódu stránky. Nejčastěji kontrolují HTML, CSS a XML, zdali dodržují pravidla daného jazyka. Existují ve formě doplňků pro webové prohlížeče, samostatných programů, či služeb. Zmíním například validátor `http://validator.w3.org`, který provozuje Konsorcium W3C, jehož hlavním úkolem je dohlížet na vývoj internetových standardů.

Vícefunkční webové nástroje

Na webu existuje řada serverů poskytujících služby užitečné pro SEO. Většinou kombinují výše zmíněné nástroje. Zpravidla stačí zadat URL stránky, o níž potřebujeme informace, a výstupem je víceméně přehledný souhrn informací. Často jsou zdarma, často obsahují kontrolu proti robotům, tzv Touringův test. Nejčastěji captchu, nebo jednoduchou otázku.

Tyto informace jsou však zpravidla jednorázové. Pro jejích dlouhodobé sledování a vyhodnocování je žádoucí, aby byly uchovávány ve formě, která umožní snadné a přehledné manipulace. Zde zbývá jen zmíněný zdlouhavý, nepohodlný a na lidském faktoru silně závislý nástroj – ruční ukládání a zpracování.

Uvádím výpis několika takových webových služeb

1. `http://www.webseoanalytics.com/` – Nástroj pro SEO analýzu, zapůsobí líbivým designem, nabízí rozšířené funkce za poplatek.
2. `http://www.seocentro.com/tools/search-engines/metatag-analyzer.html` – Zpracuje analýzu `meta` tagů, klíčových slov a podobně.
3. `http://www.seoanalyser.net` – mimo jiné odhadne hodnotu webu v dolarech a zobrazí návrhy, co by se dalo na stránce opravit.

[4]Specializovaný program, který analyzuje data z logu, a která následně prezentuje v podstatně přehlednější formě.

4. `http://www.serp.cz/` Stránka věnovaná optimalizaci SEO a SERP stránek www. Základní zjišťování pozice ve vyhledávačích na vybraný výraz do hloubky 100. pozice v 15 vybraných vyhledávačích rozdělených podle jazykové náležitosti pro český, slovenský a mezinárodní internet (jen anglicky). Rozšířený detektor pozic ve vyhledávačích (pro více výrazů) vč. srovnání pozic ve vyhledávačích s konkurencí. Nástroj SEO pro detekci site ranků stránek (Google Pagerank, Seznam S-rank, Jyxo rank, Alexa rank, zpětné odkazy na Google a Yahoo). Další funkce a nástroje.
5. `http://www.seoadministrator.cz/` Seo Administrator je placená sada SEO nástrojů pro podporu webu: *Ranking Monitor site positioning software* (více než 30 nejvýznamnějších vyhledávačů jako je Google, Yahoo, MSN, AltaVista, AllTheweb, Lycos, HotBot vč. nejvýznamnějších českých vyhledávačů, jako Seznam, Centrum a Atlas). *SEO Link popularity checker* (počet odkazů linků, které směřují na stránky, možnost vytvoření srovnávacího seznamu odkazů vašeho webového projektu a sledování jejich vývoje pomocí zobrazení nových a zaniklých odkazů). Je také schopen analyzovat stránky konkurence a vyhledávat zdroje pro potenciální výměnu odkazů. *Site indexation tool* – nástroj kontroluje indexování webových stránek. Nejprve je nutno, aby byla zaindexována vyhledávači. Ukazuje stránky webu, které již byly vyhledávači zaindexovány. Podporuje všechny hlavní a důležité světové a nejvýznamnější české vyhledáva *Link exchange tool* – soubor nástrojů pro práci s výměnnými a zpětnými odkazy. *Program Log analyzer* pro analýzu log souborů webových stránek. Srovnávací přehled využití zdrojů, počtu návštěvníků, vyhledávaných dotazů... *SEO Page Rank analyzer* – automaticky získává hodnoty Google PageRank pro list analyzovaných URL. Počet odkazů směřujících na web je také zohledněn. Dále je kontrolována přítomnost každého URL v katalozích DMOZ a Yahoo. *HTML analyzer site content analyzer* – analyzuje obsah stránky HTML. Váha a hustota klíčových slov a frází, může být také použit k analýze webů konkurence. *Program Google Data Centers* – dovoluje zaslat dotaz do všech Google's Data Centers z jednoho uživatelsky pohodlného rozhraní. *Snippets viewer snippets* – zkrácený automatický popis webu (snippets), který je zobrazován vyhledávači. Analyzuje vaše popisky (snippets) a tvoří jejich databázi.
6. `http://www.piloun.com/webinfo/` – hodnocení stránky (Rank): Google PageRank, S-rank, JyxoRank, Alexa traffic rank, Compete, validita, soubor `robots.txt`, celková velikost zdrojového textu, počet obrázků, velikost vloženého JS, počet interních odkazů, velikost vloženého CSS, počet externích odkazů, kódování, ikona stránky, počet indexovaných stránek `google.com`, `seznam.cz`, `jyxo.cz`, `bing.com`, `altavista.com`, počet odkazujících stránek, počet slov

přibližně, nejčastější slova, popis a obsah (title, description, keywords, nadpisy), měřicí ikona na stránku.

1.1.4 Hodnocení a východiska

Žádný z těchto systémů mě dokonale neuspokojil. Zejména z hlediska ceny, funkcí a účelu aplikace. Mnou vytvořený systém je zdarma a vyhovuje zadání práce.

Při volbě charakteru aplikace jsem bral v úvahu univerzálnost, přístupnost a všestrannou použitelnost, která není omezena pouze na určitý typ technologie. Těmto požadavkům nejlépe vyhovuje webová aplikace.

1.2 Prostředky a metody

Pro programování webových aplikací je v současné době k dispozici bohatá nabídka nástrojů – systémů, programů a prostředí. Např. Skripty CGI, SSL, ASP, Perl, PHP.

Ke své práci jsem se rozhodl použít následující zdroje a prostředky:

- Hardware (založený na úsporné mini-ITX desce s procesorem Intel Atom 1,6 Ghz, 1 GB RAM)
- Síťové připojení (wifi, místní poskytovatel Netlife Křižanov)
- Operační systém, distribuce: Linux, Centos 6.3
- Webový server Apache 2.2.15
- PHP 5.3.3
- Databázový server MySQL 5.1.67
- Framework Nette 2.0

K volbě těchto zdrojů a prostředků mě vedla jejich snadná dostupnost (mám je k dispozici doma a osvědčený software je použitelný zdarma a legálně). Dalším důvodem je oblíbenost a rozšířenost těchto nástrojů, což umožní zájemcům o používání či další zdokonalování a rozšiřování mého systému jeho snadnou implementaci. Důležitým důvodem je i dostupnost dokumentace a literatury a otevřená aktivní komunita uživatelů.

1.3 Instalace a nastavení serveru

Operační systém Linux Centos jsem nainstaloval standardním způsobem. Základní instalace neobsahovala webový server Apache ani databázi MySQL. Tyto součásti jsem doinstaloval pomocí nástroje YUM a nastavil konfigurační soubory. Dalším nástrojem je Nette [4]. Instalace se provádí rozbalením instalačního balíku do podadresáře webového serveru.

Nastavení serveru Apache je standardní, neobsahuje žádné mimořádnosti.

K oživení databázového serveru MySQL stačilo nastavit uživatele příkazovou řádkou MySQL. Další nastavení jsem prováděl pomocí aplikace PHPMyAdmin, kterou jsem předtím nainstaloval do podadresáře webserveru.

Framework Nette využívá aktuální verzi PHP 5.3. V podadresáři sandbox je předpřipravená struktura pro vlastní uživatelský projekt. Projektů může být více, Jedna instalace Nette dokáže obsluhovat více projektů založených na šabloně sandbox. Pro svou práci jsem se rozhodl šablonu využít a dále používat pod názvem Mimos. Pro správné fungování Nette musí nastavení PHP splňovat všechny minimální požadavky, pro jejichž kontrolu je připraven program Nette Framework Requirement Checker, jehož výstup přikládám.

Na serveru je také zprovozněno SSH na vzdálený přistup a správu. Softwarový démon `cron` pravideně spouští script `cron_scipt.php,` který naplní databázi čerstvými záznamy.

1.4 Struktura databáze

Ukládání dat obstarává databáze MySQL. Obrázek popisuje, jak jsou provázané jednotlivé tabulky. Mezi tabulkami `user` a `web` je vazba M:N (jeden uživatel může mít více webů a naopak, jeden web může být patřit více uživatelům). Tabulka `web_modules` slouží k přiřazení modulů k webu podle volby uživatele. Každý modul, který potřebuje ukládat data, má vlastní tabulku, které se využívá pro data všech sledovaných webů každého uživatele. Schéma databáze je přiloženo.

1.5 Vývoj aplikace

Postup

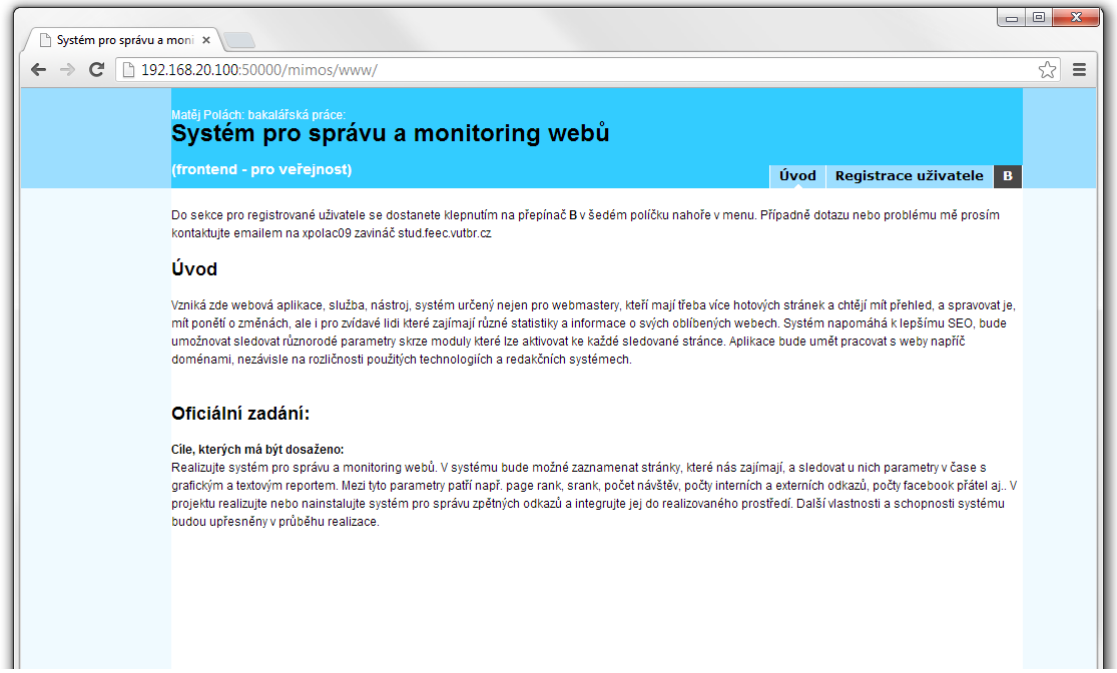
- Uživatelské rozhraní – Vytvoření uživatelského rozhraní pro FE a BE
- Role uživatelů – Vytvoření skriptů pro správu uživatelů
- Moduly – Vytvoření skriptů modulů pro sběr, analýzu a výstup dat

2 Výsledky studentské práce

2.1 Uživatelské rozhraní

Webové stránky jsou rozděleny do dvou hlavních sekci: Frontend a Backend. Mezi nimi se dá přepínat tlačítkem B/F v horní části stránky vpravo.

Frontend je veřejně přístupný a obsahuje informace a pro nové návštěvníky a nepřihlášené uživatele. Je to hlavně popis projektu, možnost registrace nového uživatele a novinky.

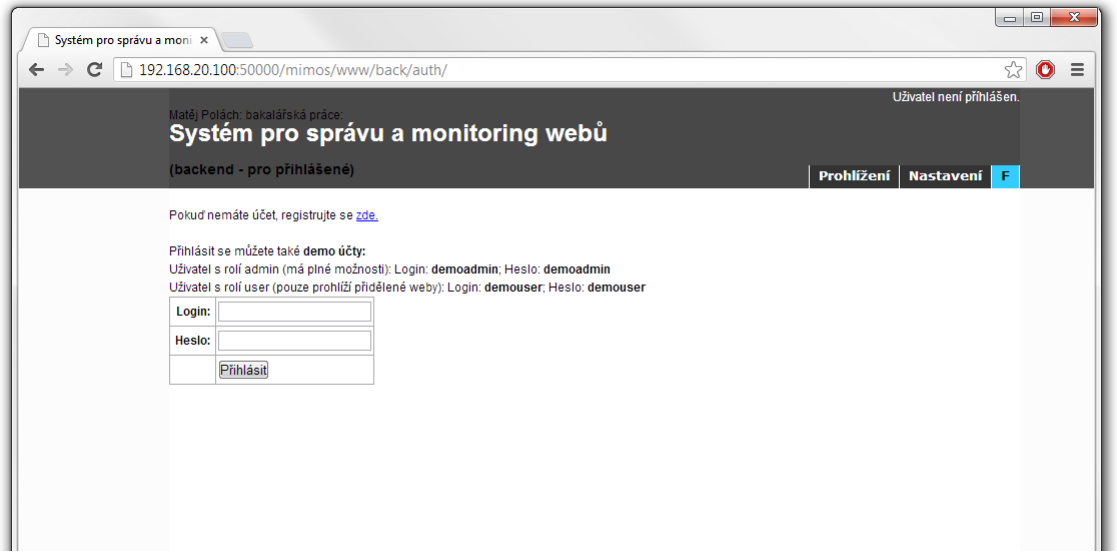


Obr. 2.1: Frontend Aplikace

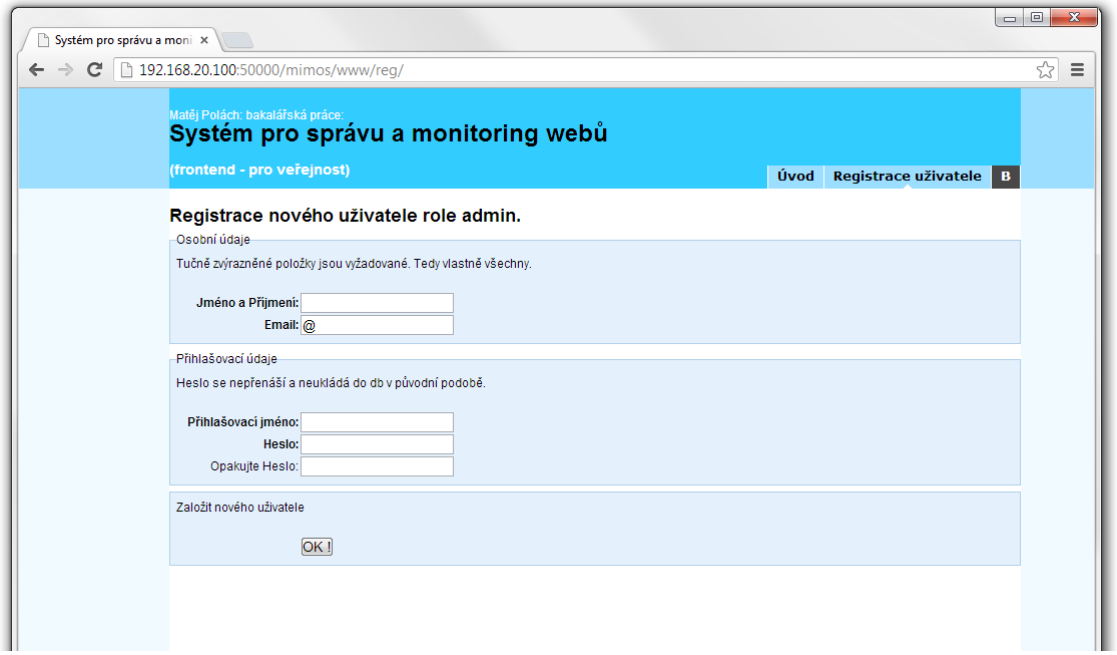
Sekce **Backend** vyžaduje přihlášení. Umožňuje přistup k hlavním funkcím systému jako je správa uživatelského seznamu sledovaných webů (založení, smazání, nastavení modulů), stránka webu s výstupem aktivovaných modulů a správu uživatelů.

2.2 ACL

Program rozlišuje dvě úrovně (role) uživatele: Admin a User. Admin má neomezený přístup. User může pouze do své sekce Prohlížení. Role přihlášeného je zobrazena v rozraní Backendu nahoře vpravo.



Obr. 2.2: Backend Aplikace



Obr. 2.3: Registrace uživatele.

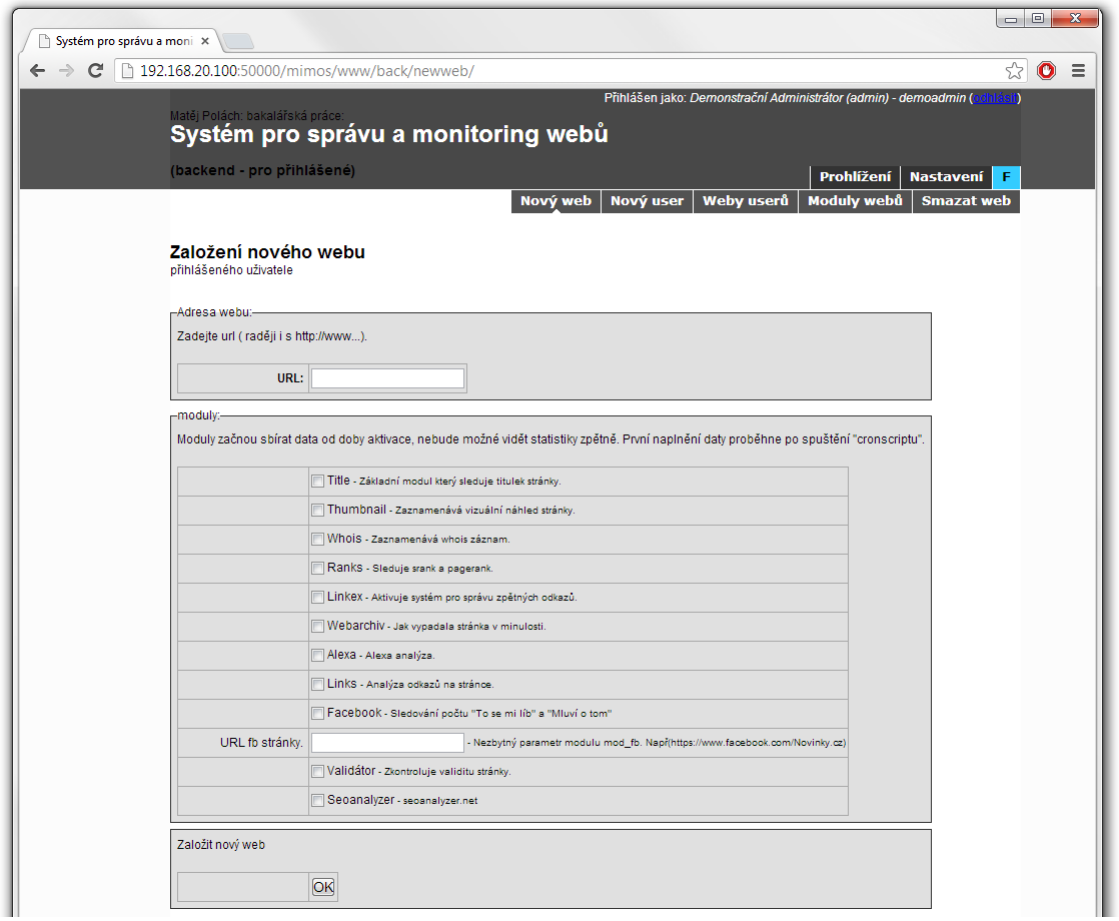
Uživatele s rolí Admin může založit kdokoli formulářem na Frontendu. Nastaví si svoje jméno a příjmení, email, přihlašovací jméno a heslo. Poté se může přihlásit (Backend) a zpřístupní se mu všechny funkce systému. V sekci Nastavení může založit své uživatele s rolí User, které tím pádem bude mít pod kontrolou.

Role User má umožněno vidět pouze seznam webů přiřazených Adminem a pro-

hlížet jejich stránky webů. Přihlašovací údaje dostane od svého Admina. Nemá přístup do žádného nastavení.

2.3 Moduly

U každého ze sledovaných webů lze aktivovat níže popsané moduly. Každý modul sleduje, zaznamenává či jinak zpracovává některý z parametrů, které se vztahují k dané webové stránce, nebo slouží jako okénko pro zobrazení rozličných služeb pracujících s webem. Stránku, ze které data získávám, většinou stáhnu funkcí PHP `file_get_contents()`. Tato data ukládám do databáze, která je popsána výše.



Obr. 2.4: Založení sledování webu s moduly.

2.3.1 Modul Title

Zaznamenává titulek stránky (obsah html tagu `TITLE`), který je nezbytnou součástí hlavičky každého dokumentu HTML. Jeho obsah se objevuje na záložce v prohlížeči,

používá se v seznamech webů (např. záložky oblíbené) a bývá nadpisem výsledků ve vyhledávačích. Výstupem modulu je tabulka s výpisem titulků v čase.

2.3.2 Modul Thumbnail

Obr. 2.5: Modul thumbnail: ukázka výpisu.

Modul Thumbnail sleduje grafický vzhled. Využívá služby `pagepeeker.com`, která na požádání vygeneruje náhled, který script následně stáhne a uloží jako obrázek do složky. Služba zdarma je vykoupena několika nedostatky: Do obrázku vkládá

razítko s logem, na první vyžání neznáme stránky vrací prázdný obrázek s animací, a využívá svoji cache – vrací stejný náhled přibližně dva týdny. Tyto nedostatky jsem potlačil. Obrázky pomocí hash otisku porovnávám a vypisuji pouze unikáty. V případě obdržení prázdného obrázku tento nezobrazuji.

2.3.3 Modul WhoIs

Pracuje s daty z databáze od registrátorů domén. Z jeho výpisu je možné zjistit vlastníka domény, datum vypršení registrace či jiné údaje.

2.3.4 Modul Ranks

Zaznamenává Google Pagerank a Seznam Srank. Jsou to čísla, která generují vyhledávače a snaží se vyjádřit důležitost či hodnotu webu. Mají vliv na pořadí ve výsledcích vyhledávání.

2.3.5 Modul Linkex

Slouží k práci s výměnnými (zpětnými) odkazy. Umožňuje jejich automatické kontrolování a správu. Tento modul se od ostatních liší. Je to externí program ve formě skriptu PHP, který jsem integroval do mého systému tak, že se v případě aktivace modulu vytvoří jeho instance k danému webu. Vyžaduje samostatnou registraci pro každý web. K dispozici je manuál na stránce `http://linkex.dk/`[5].

2.3.6 Modul Facebook

Vytáhne a zpracuje informace o počtu lídí, kteří kliknou na „líbí se to“ a lidí, kteří o tom mluví, tj. vykázali jakoukoli interakci s danou facebookovou stránkou, která má návaznost na sledovaný web v posledních 7 dnech. Tato čísla ze zobrazují na adrese např.: `https://www.facebook.com/Novinky.cz` v řádku pod úvodním obrázkem. Při aktivaci modulu je pochopitelně potřeba tuto adresu zadat.

Facebook kontroluje prohlížeč, ze kterého se na něj přistupuje. Při dolování dat jsem zjistil, že je potřeba, aby se tvářil jako Firefox. Jinak Facebook nevrátil požadovanou stránku, nýbrž pouze upozornění, že používám nepodporovaný prohlížeč. Výstup dat je realizovaný formou grafu, k jehož dynamické generaci využívám služby Google Charts.

2.3.7 Modul WebArchiv

Do systému implementuje službu `http://web.archive.org` [9]– wayback machine, která slouží pro nahlédnutí do minulosti a zobrazí starší podobu webové stránky. Služba se snaží v průběhu času archivovat webové stránky.

2.3.8 Modul Alexa

Zobrazí analýzu služby Alexa [8] daného webu. Obsahuje výpis nejdůležitějších klíčových slov z vyhledávačů, demografické složení návštěvníků a odkazy na podobné (konkurenční) stránky. Alexa sbírá údaje pomocí toolbaru v prohlížeči.

2.3.9 Modul Links

Modul Links poskytuje přehled o odkazech uvedených na sledované stránce. Stránku naparsuji pomocí DOM a získám z ní pole odkazů. Ty potom sérií podmínek testuji a třídím. Zjistím celkový počet odkazů, počet interních, externích, relativních, absolutních odkazů, počet odkazů typu `mailto`, navigačních odkazů, odkazů do subdomén a zbytek uvádím jako neidentifikované. Tato čísla se spolu s výpisem všech analyzovaných odkazů uloží do databáze a jsou využita při výpisu.

2.3.10 Modul Validator

Analyzuje stránku z hlediska validity. Používá validátor W3C [6] html.

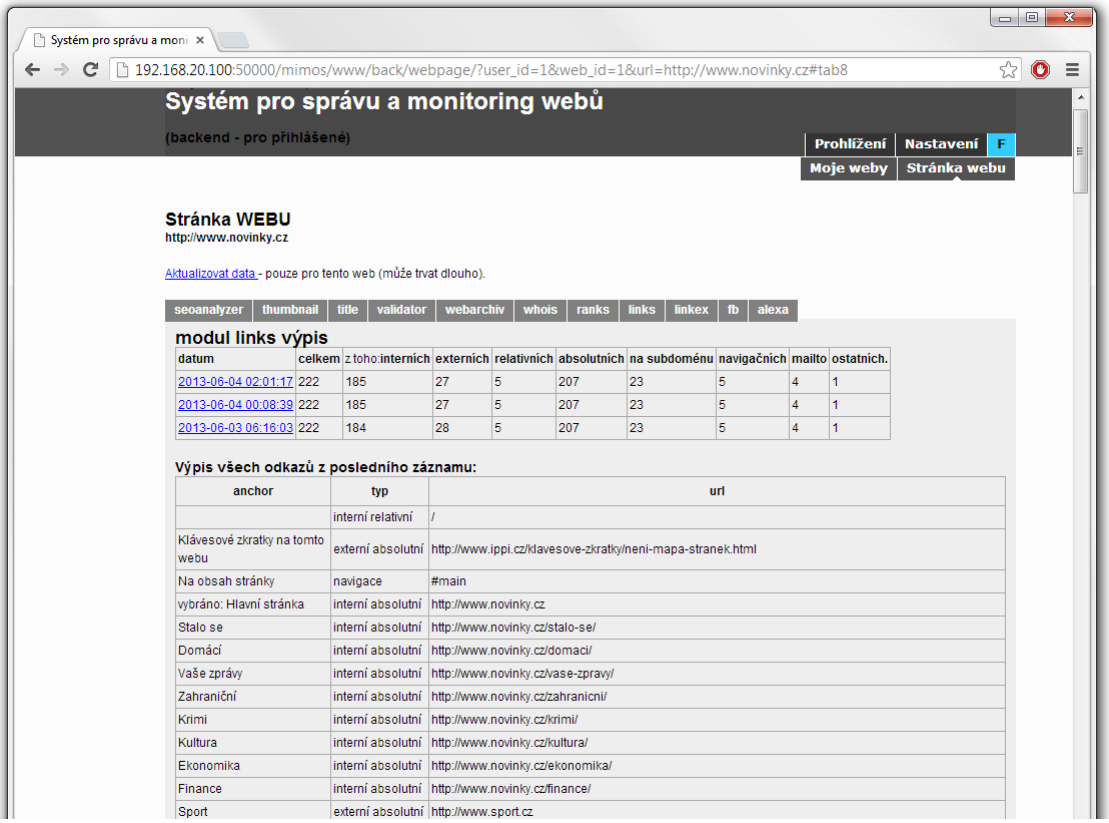
2.3.11 Modul SeoAnalyzer

Slouží k rychlému přístupu na službu `http://www.seoanalyser.net/`[7]. Zpracuje SEO analýzu, upozorní na nedostatky a poskytne doporučení k nápravě.

2.4 Testování aplikace

Testování proběhlo formou uživatelského zkoušení tak, aby byly vyzkoušeny všechny funkce, které systém nabízí. Byla provedena registrace nového uživatele `demoadmin` , heslo: `demoadmin` . Po přihlášení tento uživatel založil sledování několika webů a vytvořil uživatele `demoadmin` kterému přidělil jeden z webů.

Aktualizace dat u webu, který má aktivované všechny moduly, trvá průměrně kolem minuty v závislosti na velikosti daného webu. Vzniklé záznamy v databázi zabírají 35 kB. Obrázek náhledu stránky modulu Thumbnail ve formátu `jpeg` má kolem 50 kB. Aplikace byla testována v prohlížeči Google Chrome.



Obr. 2.6: Modul links: ukázka výpisu.

3 Závěr

Ve své bakalářské práci jsem realizoval systém pro správu a monitoring webů. Po prozkoumání literatury a existujících nástrojů v souladu se zadáním práce jsem se rozhodl pro nejvhodnější způsob řešení: Systém fungující jako webová aplikace umožňuje sledovat vizuální podobu stránky v čase formou snímání screenshotu, změny tagu `title`, Pageranku a Sranku, odkazy na stránce, jejich počet a kategorizaci, facebookové atributy „To se mi líbí“ a „Mluví o tom“. Systém umožňuje rychlý přístup k SEO analýze, validátoru kódu a výpisu WhoIs a implementuje externí nástroj pro správu zpětných odkazů LinkEX. Předností vytvořeného systému je, že funkčnost jeho modulů není závislá na přítomnosti vlastního kódu na sledovaných stránkách a umožňuje tak sledování libovolné webové stránky.

K vytvoření systému jsem použil řadu nástrojů. Webová aplikace využívá frameworku Nette, který pomáhá s generováním uživatelského rozhraní HTML a obstarává ACL (správu uživatelských účtů). Data, která získávám a zpracovávám skripty PHP, jsou ukládána do databáze MySQL. Všechny tyto nástroje zastřešuje server Apache běžící na operačním systému Linux CentOS.

Testování potvrdilo, že cíle stanovené zadáním práce byly splněny. Systém dostupný na adrese `www.polach.org` je plně funkční.

Příležitostmi ke zlepšení jsou uživatelská ergonomie a grafický vzhled. Zrychlení získávání dat by bylo možné dosáhnout eliminací opakovaného stahování téže stránky jednotlivými moduly. Praktickým omezením je přenosová kapacita domácí internetové připojky, kterou server využívá. V případě klíčového nasazení systému doporučuji zaměřit se na prvky zabezpečení. Systém je otevřený pro vkládání a upravování dalších modulů, optimalizaci a vylepšování.

Literatura

[1] Nette framework [online]. 2013 [cit. 2013-02-10]. Dostupné z URL: <http://nette.org/cs/>.

[2] Grappone, Jennifer; Couzin, Gradiva *SEO*. 2007, Zoner Press, Brno. ISBN 978-80-86815-85-5

[3] Tvorba webu [online]. 2013 [cit. 2013-06-05]. Dostupné z URL: <http://www.tvorba-webu.cz/seo/>.

[4] Nette framework [online]. 2013 [cit. 2013-06-05]. Dostupné z URL: <http://nette.org/cs/>.

[5] Dokumentace systému LinkEX [online]. 2013 [cit. 2013-06-05]. Dostupné z URL: <http://linkex.dk/documentation/>.

[6] W3C validátor [online]. 2013 [cit. 2013-06-05]. Dostupné z URL: <http://validator.w3.org/>.

[7] SEO analyser [online]. 2013 [cit. 2013-06-05]. Dostupné z URL: <http://www.seoanalyser.net/>.

[8] Alexa [online]. 2013 [cit. 2013-06-05]. Dostupné z URL: <http://www.alexa.com/>.

[9] WayBack Machine [online]. 2013 [cit. 2013-06-05]. Dostupné z URL: <http://archive.org/web/web.php>.

Seznam symbolů, veličin a zkratek

ACL Access Control List – Správa přístupových práv

CSS Cascading Style Sheet – Kaskádové styly

DOM Document Object Model – Objektový model dokumentu

HTML Hypertext Markup Language – Hypertextový značkovací jazyk pro webové stránky

JS Java Script – skriptovací jazyk používaný ve stránce HTML. Skript je spouštěný na straně prohlížeče

MySQL My Structured Query Language – Systém pro řízení databází

PHP Hypertext Preprocessor – Hypertextový preprocesor, původně Personal Home Page

SEO Search Engine Optimization – Optimalizace pro vyhledávače

SERP Search Engine Results Page – Stránka výsledků vyhledávače

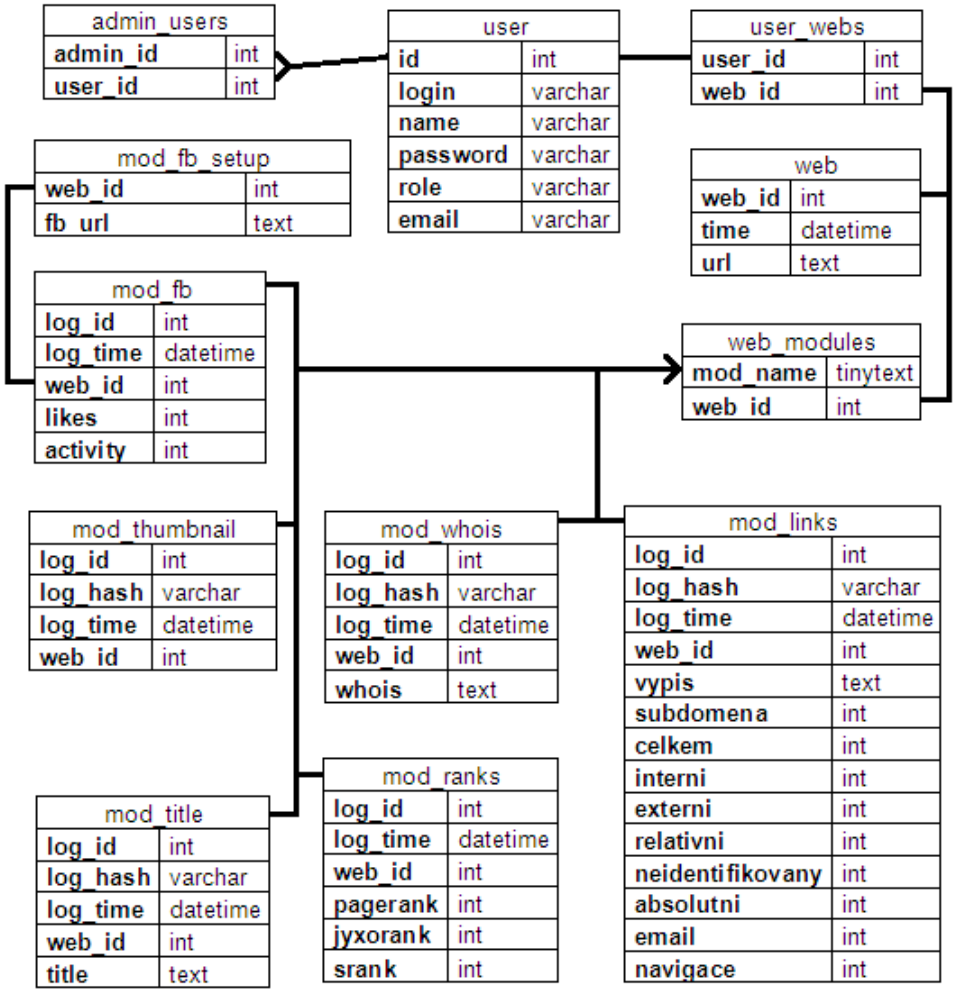
URL Uniform Resource Locator – jednoznačné určení zdroje

W3C World Wide Web Consortium – mezinárodní konsorcium zabývající se webovými standardy

–

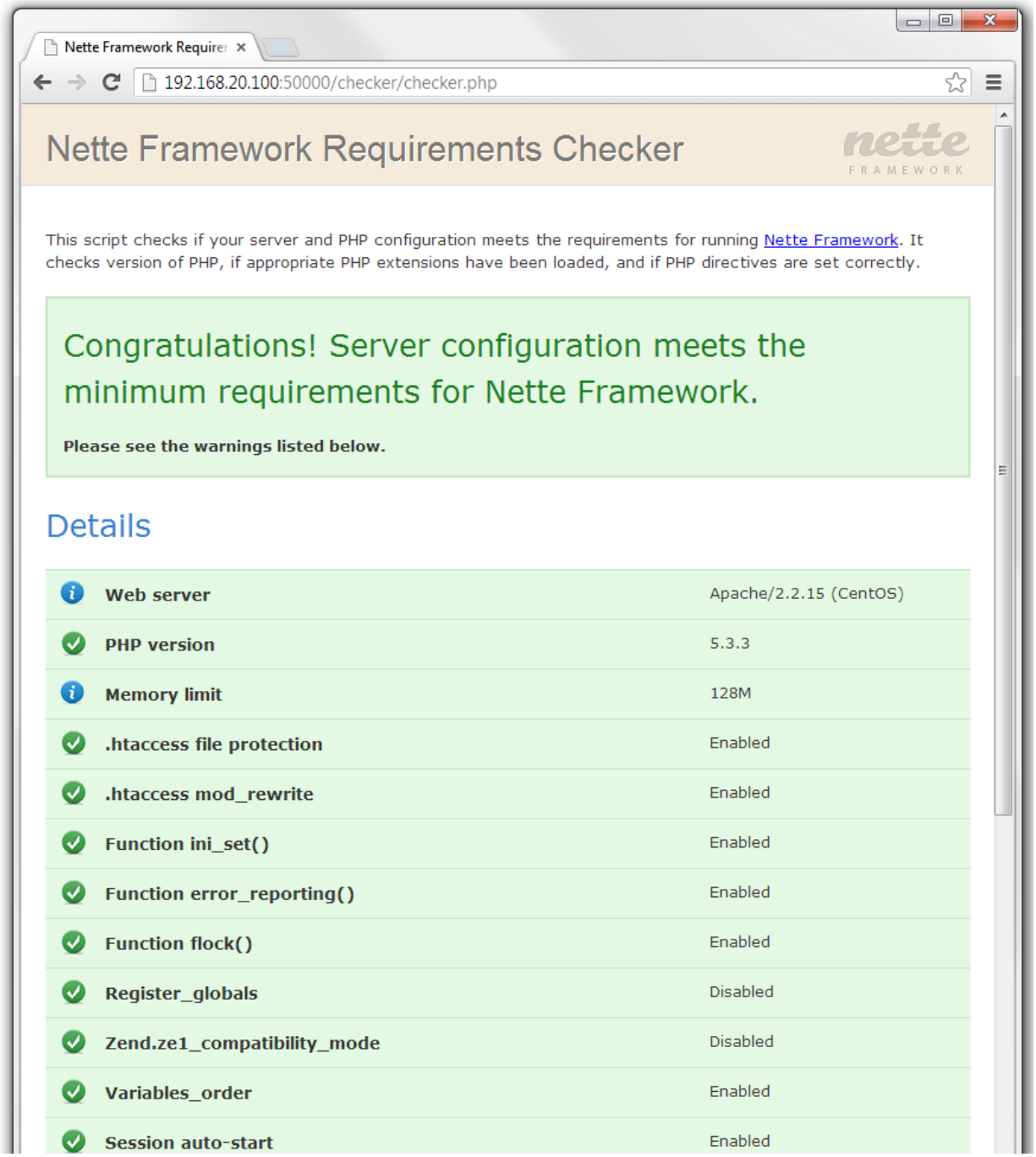
Seznam příloh

A Schéma databáze



Obr. A.1: Schéma databáze

B Nette Requirements Checker



Obr. B.1: Požadavky nette.